PREDICTING CRASH OCCURRENCE AT INTERSECTIONS IN TEXAS: AN OPPORTUNITY FOR MACHINE LEARNING

Theodore Charm
Graduate Research Assistant
Department of Government
The University of Texas at Austin
theodorecharm@utexas.edu

Haoqi Wang
Graduate Research Assistant
Department of Biomedical Engineering
The University of Texas at Austin
haoqiwang@utexas.edu

Natalia Zuniga-Garcia
Research Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
nzuniga@utexas.edu

Mostaq Ahmed
Department of Community and Regional Planning
The University of Texas at Austin
mostaq@utexas.edu

Kara M. Kockelman
(Corresponding Author)
Dewitt Greer Professor in Engineering
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
kkockelm@mail.utexas.edu

Published in *Transportation Planning and Technology* in 2023.)

ABSTRACT

This paper studies the frequency of traffic crashes at intersections across Texas by employing zero-inflated negative binomial (ZINB) models using the Maximum Likelihood Estimation (MLE) method, and various tree-based machine learning (ML) methods, namely Random Forests (RF), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and Bayesian Additive Regression Trees (BART) to predict the frequency of crashes at intersections. Official records of traffic crashes from 2010 to 2019 were used in addition to the roadway inventory database and other data sources to explore more than 700,000 intersections. Using R-square and Root Mean Square Error as metrics, results indicated that RF had the best model performance in predicting crash frequency. Resampling the data led to better prediction performances for all the models and was useful in dealing with highly imbalanced crash data. Sensitivity analysis showed that the effects of several predictors have different directions across different ML models.

Keywords: Motor vehicle crashes, intersection safety, crash counts, machine learning, imbalanced data

BACKGROUND

Traffic crashes are very expensive- they cost the society numerous human lives. Motor vehicle crashes are one of the leading killers in the U.S, with over 100 deaths every day (National Center for Statistics and Analysis 2017). In 2015, more than 2.5M Americans were taken to emergency departments due to injuries sustained in a motor vehicle crash (CDC 2018). Economically speaking, in 2017, the cost of medical care, loss of productivity, loss of lives, etc. - all sum up to more than $75 billion in the U.S (CDC 2018). Although there is a significant amount of work predicting motor vehicle crashes, there is still room for further research in order to gain a better understanding of pre-crash conditions as well as return a more accurate prediction. Those are crucial for pro-active road-safety management.

Existing literature on crash frequency prediction modeling typically adopted econometric modeling approaches (Lord and Mannering 2010; Yasmin and Eluru 2018; Wang et al. 2018). Dionne et al. (1995) and Jovanis and Chang (1986) argued in favor of the Poisson regression model in their study relating exposure variables to crash counts. In an attempt to develop crash prediction modeling for Italy, Caliendo, Guida, and Parisi (2007) investigated the comparative suitability of the Poisson, Negative Binomial, and Negative Multinomial distributions. They found that when there was over-dispersion in the crash data, the Poisson model would have weaker predictive power than the negative binomial model. Another problem with crash data is the presence of a lot of zeroes in the dataset, where the zeroes indicate the locations in which no crashes occurred. Studies found that due to unobserved heterogeneity and the presence of excess zeros, the ZINB model performed better than the regular negative binomial model (Greene 2007; Dong et al. 2014). Nonetheless, these econometric models often fail to make accurate predictions when working with complex and highly nonlinear motor vehicle crash data (Karlaftis and Vlahogianni 2011). To deal with the limitations of statistical models, several ML techniques, including decision tree-based models, Artificial Neural Network (ANN), Support Vector Machine, and deep learning models, have been applied to various traffic crash prediction models (Chong, Abraham, and Paprzycki 2005; Cho et al. 2014). That is because ML models do not rely heavily upon certain types of underlying assumptions when examining the relationships between the dependent variable and the contributing factors (Dong et al. 2018; Rahman et al. 2019). Among many ML techniques, tree-based models are being widely used in traffic safety literature because of their capability to identify the complex pattern of crash likelihood and their interpretability in explaining the relationship between target variables and the predictor variables (Chang and Chen 2005; Rahman, Kockelman, and Perrine 2022; Zuniga-Garcia, Perrine, and Kockelman 2022). In another study, Liu, Chen, and Yang (2008) compared the prediction accuracy of the negative binomial regression model with ANN in crash frequency prediction and found that ANN offers higher accuracy relative to the negative binomial model. Dong et al. (2018) developed a deep learning model with a multivariate negative binomial regression layer and concluded that the model provides better traffic crash prediction across different levels of injury severity.

To address the heterogeneity of the crash data, some studies applied data clustering methods prior to applying ML models for crash prediction and examined the effectiveness of clustering treatment for most cases (de Oña et al. 2013; Eluru et al. 2012; Kaplan and Prato, 2013; Zhao, Iranitalab, and Khattak 2019). Many previous studies used accuracy or a loss function optimized for accuracy as the validation tool to measure the performance of the crash prediction model (Abdelwahab and Abdel-Aty 2001; Yu and Abdel-Aty 2013; Kingma and Ba 2017; Zheng et al. 2019). The problem with using prediction accuracy as the only metric is that it can be misleading due to the highly imbalanced traffic crash data (Rahim and Hassan 2021; Guo et al. 2008). Accuracy puts higher weight on the common class in an imbalanced dataset which leads to poor performance for rare classes like fatal crashes. A number of recent studies include precision and recall metrics to deal with the imbalanced data problem which penalizes the model for discounting the rare classes (Elamrani Abou Elassad, Mousannif, and Al Moatassime 2020; Fiorentini and Losa 2020). A high precision and recall value for a class implies the model made good classification predictions, whereas a low value implies poor classification predictions.

In the traffic safety literature, prediction accuracy was not the main focus of the statistical models; the main focus was to use the models to investigate the contributing factors of crash events and different levels of crash severity (Iranitalab and Khattak 2017; Rahim and Hassan 2021). The prediction performance was used primarily for validation purposes in statistical models. On the other hand, ML models are mostly employed as prediction tools in the traffic safety literature with higher accuracy but less interpretability than statistical models.

In this era of ML and deep learning, many cutting-edge techniques are still underexplored in the study of motor vehicle crashes. Moreover, few studies considered land-use and demographic variables in predicting crash frequency. Most importantly, the majority of the crash prediction literature focused on road segments, whereas crashes occurring at intersections received relatively little attention (Zuniga-Garcia, Perrine, and Kockelman 2022). That said, a significant proportion of motor vehicle crashes occurred at intersections. Among the 5.63M crashes recorded on public roads across the state of Texas from 2010 to 2019, approximately 20% of them occurred at intersections. Given intersections generally have more complex geometry, they are very important from a traffic safety perspective. In light of the gap in literature, this paper aims to contribute to the study of motor vehicle crashes by devising an innovative approach to predicting crash occurrence at intersections in Texas, as well as examining the contributing factors through comparing the predictions of various econometric and ML methods. Since over 70% of the intersections had 0 crashes recorded, this paper used a ZINB model estimated by MLE. A series of tree-based ML methods, namely RF, XGBoost, LightGBM, and BART, were used to predict the frequency of crashes at intersections. To handle the highly imbalanced crash data, the dataset has been resampled by implementing the *ovun.sample* function of the *ROSE* package in R, which is a "bootstrap-based technique that helps the task of binary classification in the presence of rare classes". The empirical results identify the key predictor variables for motor vehicle crashes. They suggest vital policy implications and hold promise for a safer transportation system nationwide.

DATA

Crash records from 2010 to 2019 were acquired from the Texas Department of Transportation (TxDOT) Crash Records Information System or "CRIS" (C.R.I.S. 2020). The CRIS system collects crash reports occurring on public roadways across all 254 Texas counties, as recorded by the police. To appreciate network-level design and use information, this paper also acquired data from the TxDOT Roadway Inventory database.

The CRIS crash records were spatially matched with local land use, several census-tract level variables including population, employment, median household income, median age (Khattak et al. 2002), and precipitation, i.e., snow and rain (Khattak, Kantor, and Council 1998), as well as other details (like distances to the nearest hospital or school, and transit stop density). Specifically, the crash records were spatially matched to the nearest census tract centroid. The census tract-level variables were obtained from the American Community Survey dataset (ACS 2020). The 2015-2019 ACS 5-year estimates were used in the analysis. This paper also used annual rainfall data (1981 to 2010) from the Texas Water Development Board (2014) to obtain county-level average yearly precipitation.

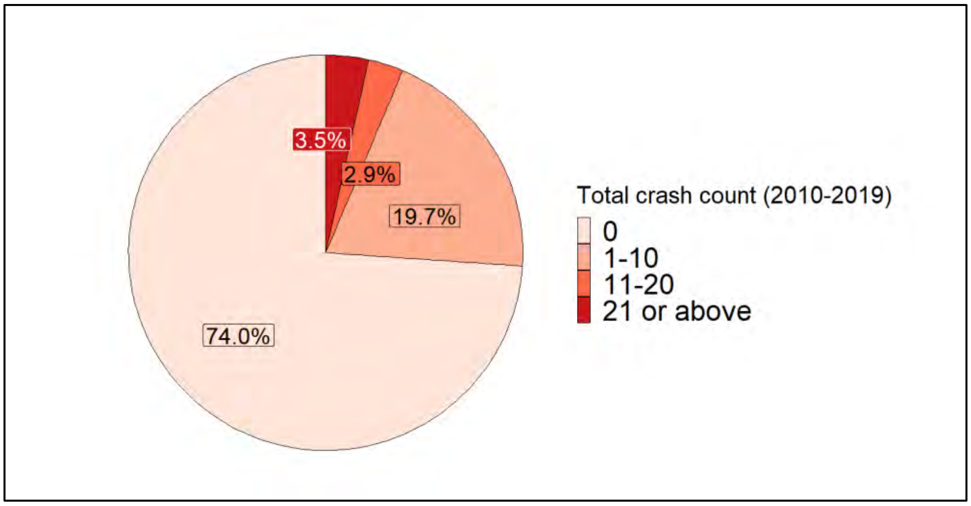


Figure 1: Crash counts per Texas intersection in 2010-2019 (n = 707,161 intersections)

Total crash counts for each Texas intersection over the recent 10-year period were obtained. Among all intersections, 522,933 (74%) had 0 crashes recorded over the 10-year period, 19.7% had 1 to 10 crashes, 2.9% had 11 to 20 crashes, and fewer than 4% had 21 or more crashes. The mean crash count was 3.18 per intersection. Figure 1 illustrates the distribution of the crash counts per intersection, and Table 1 provides summary statistics of the variables at the intersection and census-tract levels.

Table 1: Summary Statistics for Intersection Crash Count Model Variables

Variable	Mean	Std. Dev	Min	Median	Max
Total police-recorded crashes from 2010 to 2019	3.18	15.62	0	0	996
Length of sidewalk within 150 ft of intersection centroid	10.81	63.72	0	0	1092
Number of lanes major approach[1]	2.23	0.72	1	2	8
Number of lanes minor approach	2.03	0.25	0	2	8
Presence of median on the major approach	0.014	0.12	0	0	1
Presence of median on the minor approach	0.0021	0.046	0	0	1
Intersections located on the TxDOT system	0.16	2.14	0	0	1
Median width major approach (ft)	0.56	7.70	0	0	519
Median width minor approach (ft)	0.085	3.35	0	0	519
Lane width major approach (ft)	10.5	2.11	0	10	49
Lane width minor approach (ft)	9.85	1.26	0	10	49
Shoulder width major approach (ft)	0.72	2.34	0	0	38
Shoulder width minor approach (ft)	0.065	0.70	0	0	32
Annual average daily traffic (AADT) major approach	1,141	3,208	0	188	142,733

[1] The vast majority of the intersections have 2 lanes at the major approach. 2-lane approach constitute 89.1% of the intersections, followed by 4-lane (8.7%), 3-lane (0.6%), and 1-lane (0.1%).

Annual average daily traffic (AADT) minor approach	221	607	0	136	62,054
Percentage of truck in the major approach	4.85	5.43	0	3.2	95.8
Percentage of truck in the minor approach	3.44	2.25	0	3.2	93.3
Walk-miles traveled per area[2]	326	454	0	155	15,339
Walk-miles traveled per capita	0.14	0.035	0.094	0.13	0.40
Walk-miles traveled	772	484	0	675	4,443
Speed limit major approach (mph)	57.02	6.50	10	58.88	85
Speed limit minor approach (mph)	58.54	3.03	10	58.88	85
Local major approach	0.67	0.47	0	1	1
Local minor approach	0.93	0.25	0	1	1
Collector major approach	0.18	0.38	0	0	1
Collector minor approach	0.052	0.22	0	0	1
Arterial major approach	0.14	0.12	0	0	1
Arterial minor approach	0.015	0.12	0	0	1
Unknown major approach	0.0067	0.082	0	0	1
Unknown minor approach	0.00090	0.030	0	0	1
Rural (pop: <5,000)	0.27	0.44	0	0	1
Small urban (pop: 5,000-49,999)	0.12	0.32	0	0	1
Urbanized (pop: 50,000-199,999)	0.11	0.31	0	0	1
Large urbanized (pop: 200,000+)	0.50	0.50	0	0	1
Signalized intersection[3]	0.02	0.15	0	0	1
Number of approaches arriving in the intersection	3.19	0.68	0	3	5
Distance to nearest school (miles)	1.41	2.28	0	0.55	18.64
Distance to nearest hospital (miles)	5.10	5.16	0.017	2.83	18.64
Transit presence within 0.25 miles of intersection centroid	0.021	0.14	0	0	1
Count of transit stops within 0.25 miles of intersection centroid	0.067	0.62	0	0	26
Population density (per acre)	3.51	3.92	0	2.18	96
Job density (per acre)[4]	2.71	3.07	0	1.62	65.66
Median income (in USD)[5]	32,370	13,792	2,499	29,025	124,355

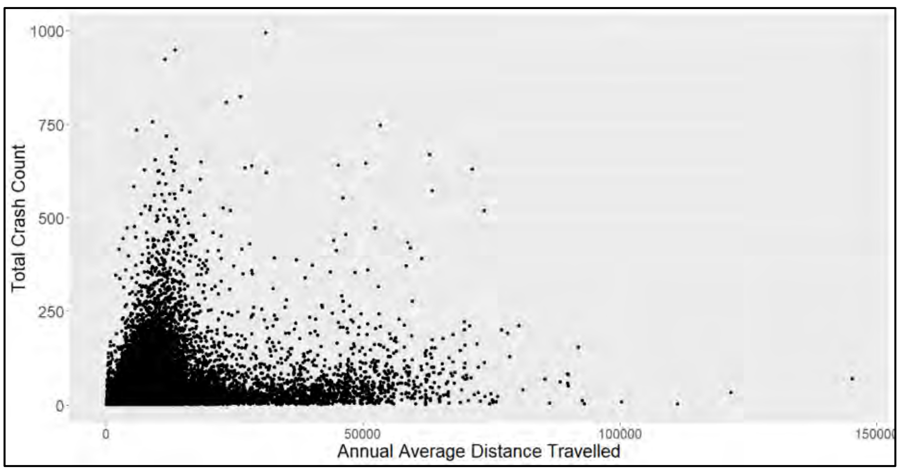
[2] Walk-miles traveled was obtained via responses to the 2016/2017 National Household Travel Survey.

[3] Signalized intersections constitute merely 2.2% of the intersections.

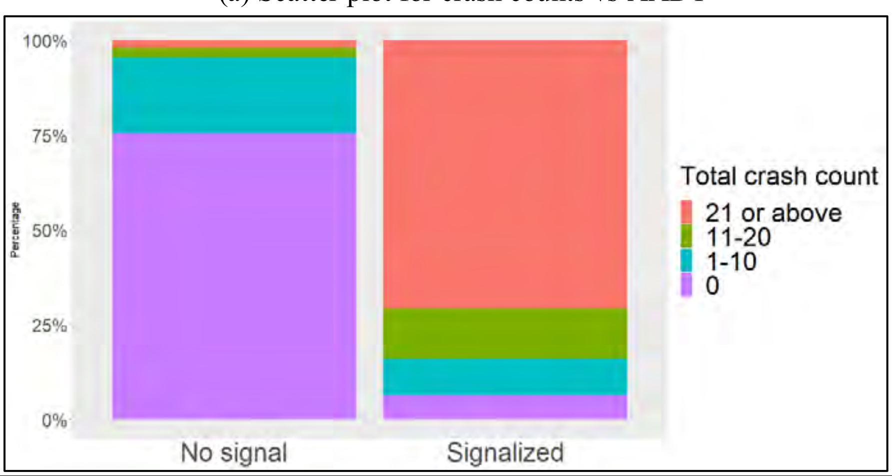
[4] Population and employment densities were calculated by dividing the total population (or jobs) by the areas (in acres) of each census tract, using the 2015-2019 ACS 5-year estimate.

Median age[6]	37.25	6.72	18.8	36.5	73.7
Average yearly precipitation (1981 to 2010) (inches)[7]	36.62	11.18	9.85	37	59.59

The association between crash counts and a number of explanatory variables is illustrated in Figure 2, in particular, annual average daily traffic (AADT), signalized intersection, and number of lanes at major approach. The sum of AADTs for the major and minor approaches was computed, and the crash counts against the sum of AADTs is plotted in Figure 2a, which shows that intersections with frequent crashes tend to have higher-than-average AADTs. Figure 2b shows that most intersections with no signals had very few crashes. Nonetheless, a high proportion of signalized intersections had relatively high numbers of crashes. Specifically, about 70% and 40% of signalized intersections had more than 20 crashes and 50 crashes from 2010 to 2019, respectively. Figure 2c illustrates that with the increase of number of lanes crash counts at the intersections increase. For example, most intersections with 1 or 2 major lanes had no crashes, but about 37% of intersections with 5 to 6 major lanes had 21 or more crashes. About 40% of intersections with 7 to 8 major lanes had over 50 crashes. Figure 2 provides evidence that intersection crashes are positively correlated with AADT, signalized intersections, and number of lanes at major approaches.



(a) Scatter plot for crash counts vs AADT

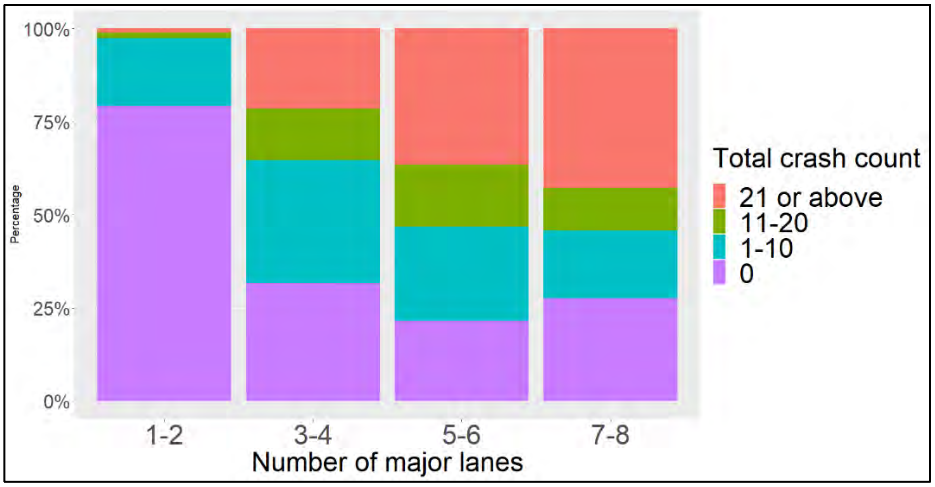


[5] This was measured by the median household income of each census tract, using the 2015-2019 ACS 5-year estimate.

[6] This was measured by the median age of each census tract, using the 2015-2019 ACS 5-year estimate.

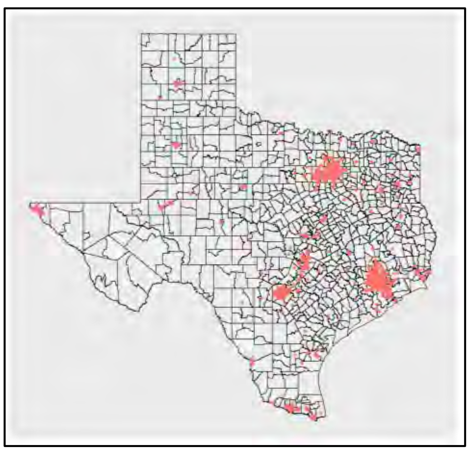
[7] This was measured by the average yearly precipitation of each county, from 1981 to 2010, using the Texas Water Development Board precipitation data.

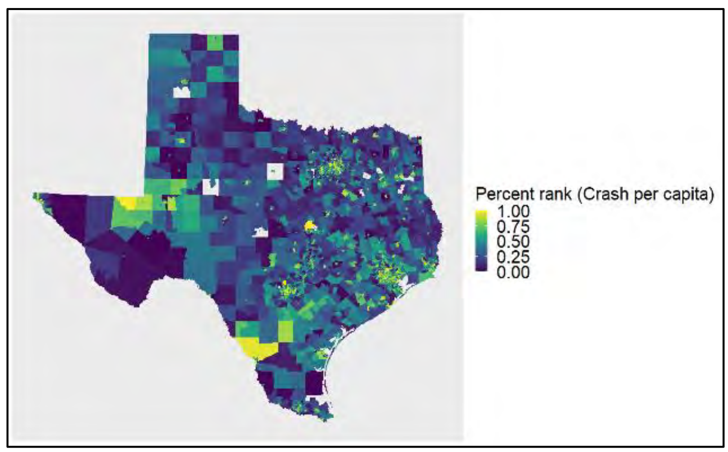
(b) Percentage of intersections by crash count range vs signalized and unsignalized intersections



(c) Percentage of intersections by crash count range vs lane count

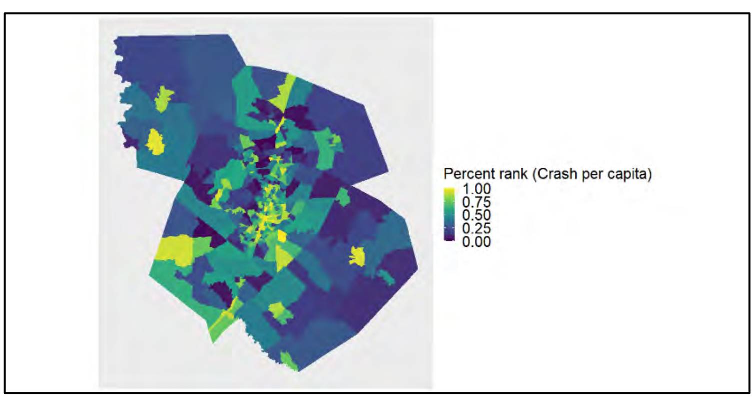
Figure 2. Crash counts by AADT, presence of traffic signal and number of lanes





(a) Texas intersections where # crashes ≥ 100 over 10-year period

(b) Crashes per capita across 5,265 Census tracts



(c) Crashes per capita in 6-county Austin region

Figure 3: High-crash intersections (a) and Census-tract level crash rates (per capita) (b & c)

This paper also provides visualization of crash counts at the census-tract level. Figure 3a locates the intersections that had more than 100 crashes over the 2010-2019 period (i.e., more than 10 crashes per year on average). Such intersections are represented by the red dots in the map. Figure 3b illustrates the number of crashes per capita for each census tract. Crashes per capita were computed by dividing the total number of crashes by the population within the census tract. The percent ranks for each value are represented by different colors. Higher percent ranks are closer to the yellow end of the color spectrum, while lower percent ranks are closer to the purple end. The yellow spots are concentrated in large cities in Texas, where the census tracts have higher population densities. That indicates that large cities are more likely to have higher average crash counts than their rural counterparts. Specifically, this suggests there is an association between population density and crash counts at the census tract level. To better capture this relationship, this paper examines the Austin metropolitan area as an example, which includes the counties of Bastrop, Burnet, Caldwell, Hays, Travis, and Williamson. Figure 3c illustrates that the more densely populated census tracts tend to have higher average crash counts, in particular Travis County and the areas along the IH-35 corridor. In the next section, statistical models were used to study the association of crash counts with the explanatory variables.

REGRESSION MODEL

As a baseline, ZINB models were calibrated to appreciate the effects of various explanatory variables on the total (10-year) crash counts at each of Texas' 707,161 intersections. Since over 70% of intersections had 0 crashes recorded, this paper used a zero-inflated count model. As the standard deviation of the outcome was significantly higher than the mean, the data was over-dispersed. In light of this issue, this paper used a negative binomial model instead of a Poisson model. As a result, ZINB models were employed for regression analysis. This paper included all explanatory variables in the ZINB regression. Since the variables were measured on different scales, this paper standardized all explanatory variables to make the values comparable. First, the means were subtracted from the original values. Second, the resulting values were divided by the standard deviation to acquire the standardized values.

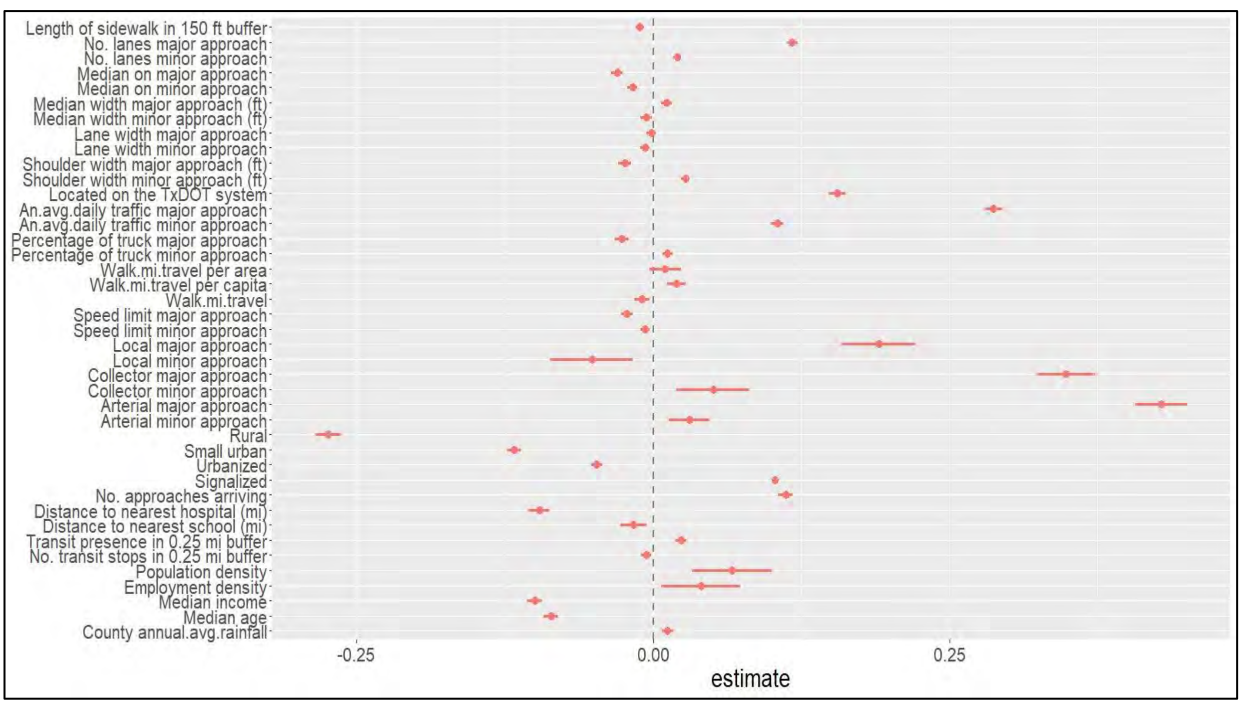


Figure 4. Coefficient plot of ZINB model

Figure 4 presents the coefficient plot for the ZINB model. One may interpret the coefficients as follows: "For one unit change in the predictor, the difference in the logs of expected counts of the outcome variable is expected to change by the respective regression coefficient, given other predictors are held constant" (UCLA 2022). The results in Figure 4 are to a large extent consistent with Figure 2. Crash counts are positively correlated with AADT, signalized intersections, and number of lanes in the major approach. Population density also has a positive effect on traffic crashes. Regarding other road-specific attributes, presence of median on approaches and lane widths of approaches show negative effects. Walk-miles traveled per capita increases crash counts, while speed limits at the approaches tend to decrease crash counts. Number of approaches arriving in the intersection has a positive effect. Other location features also show significant effects. Increasing distance to the nearest hospital reduces crash counts, while the presence of transit within 0.25 miles of the intersection centroid increases crash counts. As for census-level attributes, population density and average annual rainfall demonstrate positive effects, whereas median income and median age show negative effects.

TREE-BASED ML MODELS

Next, various tree-based ensemble ML models were used to predict crash occurrences at intersections across Texas, including RF, XGBoost, LightGBM, and BART. The models had 42 features in total. This paper evaluated the performance of the models in the predictions. The procedures were as follows: (1) randomly split the data into 70% training and 30% test sets; (2) fit the model on the training data and generate predictions; and (3) evaluate model performance with various metrics, namely R-square and root mean squared error (RMSE).

RF

A RF regression constructs decision trees for training. Depending on the size of the training set and predictions of individual decision trees, the RF algorithm determines the number of decision trees used (Greenwell and Boehmke 2020). Specifically, the decision trees are generated by "splitting each node using the best among a subset of predictors randomly chosen at that node with a different bootstrap sample of the data" (Zhao et al. 2021). The RF method computes the final prediction value based on the average prediction of individual decision trees (Liaw and Wiener 2002). For the hyperparameter tuning, the number of trees was set to 500 in the RF regression. This paper used the squared error to measure the quality of the split and considered all features when looking for the best split.

XGBoost

Chen and Guestrin (2016) devised the XGBoost method as a scalable ML system for gradient tree boosting. XGBoost constructs consecutive small trees with each tree correcting the net error from the previous trees (Chen and Guestrin 2016). XGBoost is trained in a forward "stage-wise" manner, aiming to minimize the sum of squared errors by tuning the parameters continuously (Li and Kockelman 2022). "The first tree is split on the most predictive feature, and then the weights are updated to ensure that the subsequent tree splits on whichever feature allows it to correctly classify the data points that were misclassified in the initial tree. The next tree will then focus on correctly classifying errors from that tree, and so on. The final prediction is the weighted sum of all individual predictions" (Zhao et al. 2021). As to hyperparameter tuning, the maximum depth of the trees was set to 6, the number of rounds for boosting was 500, and learning rate was 0.1 in the XGBoost training model.

LightGBM

The LightGBM method incorporates gradient-based one-side sampling (GOSS) and exclusive feature bundling, and it is particularly useful for large datasets (Ke et al. 2017). The GOSS algorithm keeps all the instances with larger gradients while randomly dropping those instances with smaller gradients (Li and Kockelman 2022). LightGBM speeds up the training process, thus reducing the computational time significantly. In the LightGBM model, the leaves per tree was set to 6, number of threads was 2, number of boosting iterations was 1000, and learning rate was 0.1.

BART

BART is a Bayesian non-parametric approach that fits a model using an influential prior distribution (Chipman, George, and McCulloch 2010). BART is a Bayesian "sum-of-tree" model in which "each tree is constrained by a regularization prior to be a weak learner" (Chipman, George, and McCulloch 2010). It performs iterative fitting and inference through conducting the back-fitting Monte Carlo Markov Chain that generates samples from a posterior. BART is robust to hyperparameter settings and addresses uncertainties with a Bayesian approach (Zhao et al. 2021). However, the method requires a lot of memory and time for computation. The number of trees was set to 100 for the model's training.

COMPARISION OF MODEL PERFORMANCE

Balanced and Unbalanced Data

In the crash dataset, there were 522,933 (74%) zero-count intersections and 184,228 (26%) non-zero-count intersections. That made the data highly imbalanced. To address this issue, the dataset was resampled by implementing the *ovun.sample* function of the *ROSE* package in R, which is a "bootstrap-based technique that helps the task of binary classification in the presence of rare classes" (Lunardon, Menardi, and Torelli 2014). *ovun.sample* generated synthetic balanced samples through a combination of randomly oversampling the minority class (intersections with non-zero crashes) and undersampling the majority class (intersections with zero crashes). In particular, it used bootstrapping to draw synthetic samples from the feature space neighborhood around the minority class to create new rows of new data for the minority class. It also randomly selected a set of majority class observations and removed those observations from the dataset (He and Garcia 2009). After resampling, the numbers of zero-crash and non-zero crash intersections were approximately equal (zero crash: 353,813 and non-zero crash: 353,113), thus the balance of the dataset was adjusted. The modified sample was denoted as balanced data.

Signalized vs Unsignalized Intersections

Signalized intersections and AADTs exerted disproportionately high weights on the model predictions (as shown in Figure 5). As a result, other features were not well accounted for in the predictions. To deal with this problem, this paper subsetted the data into signalized intersections and unsignalized intersections. It also only included the intersections where the sum of AADTs of the incoming links exceeded 500 (i.e., excluding the low-volume sites). After subsetting the data, this paper found that there were 15,222 signalized intersections and 235,822 unsignalized intersections. Among the unsignalized intersections, 121,983 had zero crashes and 113,839 had non-zero crashes.

R-square and RMSE

Table 2 presents the summary of the model performances, in terms of R-square and RMSE. R-square and RMSE are commonly used metrics to evaluate model fit and performances for ML models (Li and Kockelman 2022). Using the original (or imbalanced) data, we found that the R-squares of the ML models were not particularly high. The ZINB model produced the worst predictions, as it yielded the lowest R-square and highest RMSE among all models. Concerning the four ML models, RF regression resulted in the highest R-square (0.534) and BART had the lowest R-square (0.508). The RMSE ranged from 10.64 to 19.71. LightGBM yielded the lowest RMSE, followed by RF. The RMSEs indicated unsatisfactory predictions of the models. There were two possible reasons for this issue. First, the data contained a high proportion of zero-crash intersections. Second, there were a number of extreme values. For example, the maximum number of crashes was 996. Model predictions are likely to be affected by the extreme values. Resampling the data led to better predictions for some of the models. The R-squares increased across the models, with RF reaching a R-square of above 0.8. RMSEs, on the other hand, only showed improvement for three models. RMSEs for the RF, XGBoost, and ZINB models decreased by 2.07, 0.19, and 7.36, respectively. It is evident that after resampling, RF's RMSE saw the most significant improvement. Nonetheless, the RMSEs for LightGBM and BART increased by 1.92 and 7.82, respectively, which indicated poorer predictions for the two models, especially BART. As to the computation times, ZINB was the fastest model, while BART took the longest time to compute (712 minutes), followed by RF (508 minutes).

Table 2: Comparison of model performance: Imbalanced vs balanced data

	Imbalanced data (N=707,161)		**Balanced data (N=706,926)**		**Comp. time (min)**
	R-square	**RMSE**	**R-square**	**RMSE**	
ZINB	-1.442	59.03	-5.979	51.67	14
RF	0.534	10.66	0.832	8.59	508
XGBoost	0.527	10.69	0.753	10.50	84
LightGBM	0.531	10.64	0.647	12.56	19
BART	0.508	19.71	0.602	27.53	712

Table 3 compares the performances of the ML models between signalized and unsignalized intersections. It shows that the R-squares are comparable across the two groups, but the RMSEs are much higher for signalized intersections. This is partly due to the higher variation of crash counts, in particular the higher number of extreme values, at signalized intersections. Unsignalized intersections had many more zero crash counts, thus yielding lower RMSEs that are comparable to Table 2. Considering model performances, one can see that the RF model yielded the best model performance overall.

Table 3: Comparison of model performance: Signalized vs Unsignalized intersections

	Signalized (N=15,222)		**Unsignalized (N=235,822)**	
	R-square	**RMSE**	**R-square**	**RMSE**
RF	0.245	63.12	0.287	10.47
XGBoost	0.243	63.66	0.241	10.67
LightGBM	0.261	62.87	0.232	10.74
BART	0.203	83.69	0.194	14.27

Feature Importance

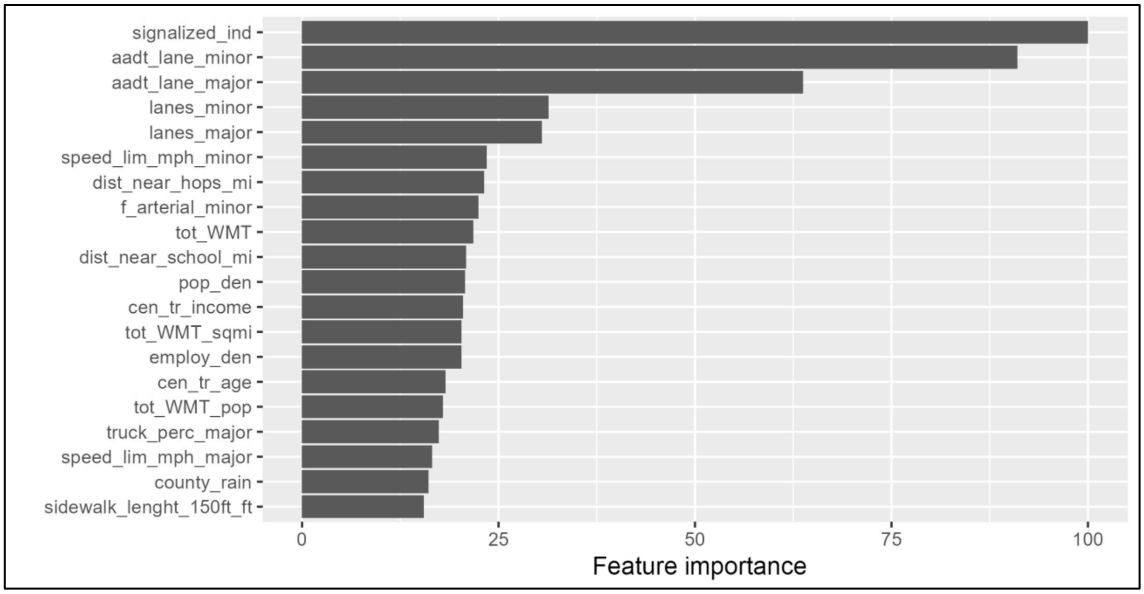
Given that RF had the best model predictions in the analysis, this paper is interested in feature importance, that is, the relative importance each feature has on the predictions of the RF model

(Casalicchio, Molnar, and Bischl 2018). This paper calculates the model-specific feature importance scores for RF. The importance scores are computed through permuting out-of-bag (OOB) data to obtain validation-set errors for individual decision trees[8]. Each predictor variable is then randomly permuted in the OOB data and the error is calculated again. The difference between the two errors is obtained for the OOB data and subsequently averaged over all trees in the forest (Greenwell and Boehmke 2020). If a predictor X is important, then a change in X's value in the OOB data will contribute to a larger increase in the validation error compared to other predictors (van der Laan 2006).

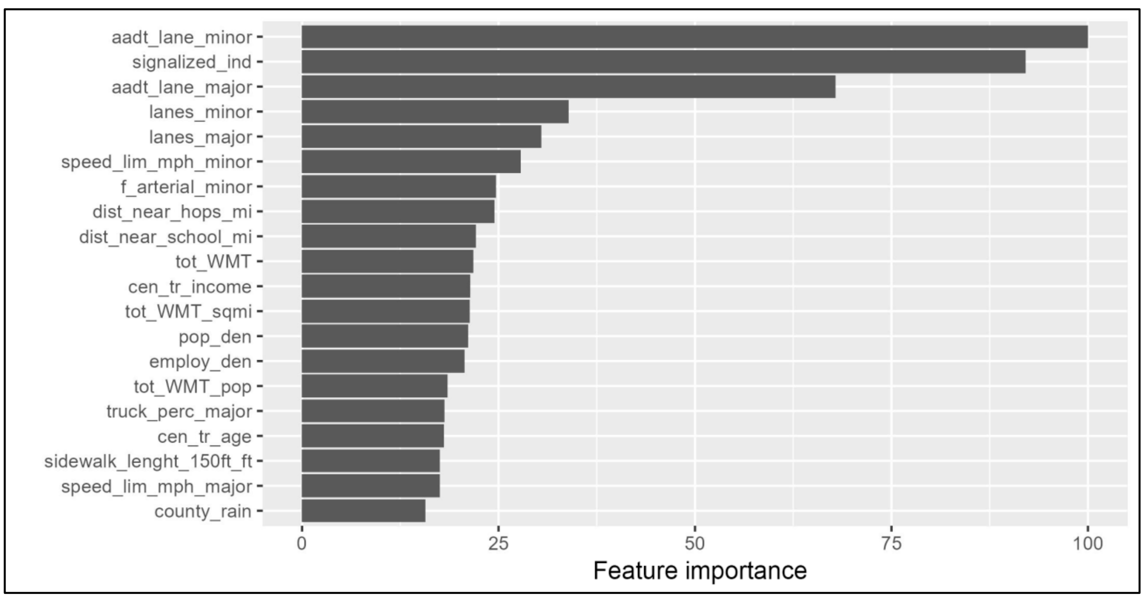
This paper employed the *vip* package in R to calculate feature importance, and it evaluated the top 20 features in terms of importance (Greenwell and Boehmke 2020). It scaled all measures of importance, such that the top feature had a maximum value of 100. Figures 5a and 5b illustrate the feature importance of individual features using the imbalanced and balanced data, respectively. The figures show that signalized intersections are the top feature, followed by AADTs, number of lanes of the approaches, and speed limit of the minor approach. Other important features included distance to the nearest hospital, distance to the nearest school, arterial minor approach, and walk-miles traveled. A number of census-tract level attributes were also important, including population density and median income. It is noteworthy that signalized intersections, and to some extent AADT at minor approach and AADT at major approach, had exceptionally high feature importance compared to other features. The three features exerted disproportionately high weights on the RF model predictions.

As explained in previous section, the data was subsetted to focus on features other than signalized intersections and AADTs. Analyzing only the high-volume intersections where the sum of AADTs exceeded 500, Figures 6a and 6b illustrate the feature importance for signalized and unsignalized intersections, respectively. They found that total walk-miles traveled, distance to the nearest school, distance to the nearest hospital, population density, and employment density were the most important features to the model predictions, although the five features were ranked differently between signalized and unsignalized intersections.

[8] When conducting bootstrap aggregating, two datasets are generated, namely the bootstrap sample and OOB set. While bootstrap sample is selected to be "in-the-bag", OOB set is all data that are not selected in the sampling process (James et al. 2013).

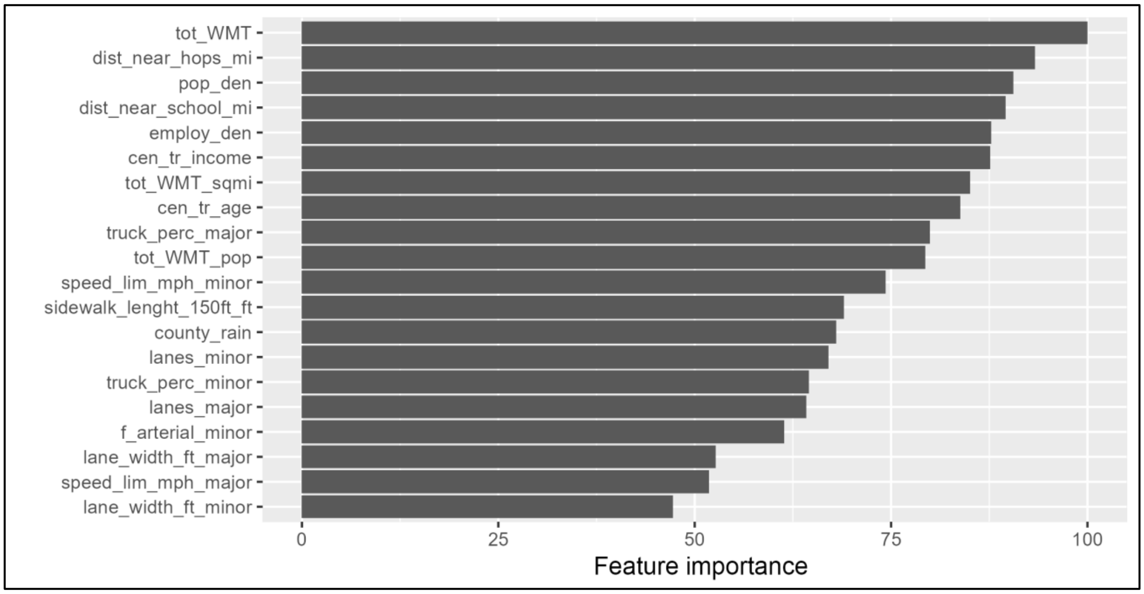


(a) Feature importance for unbalanced data

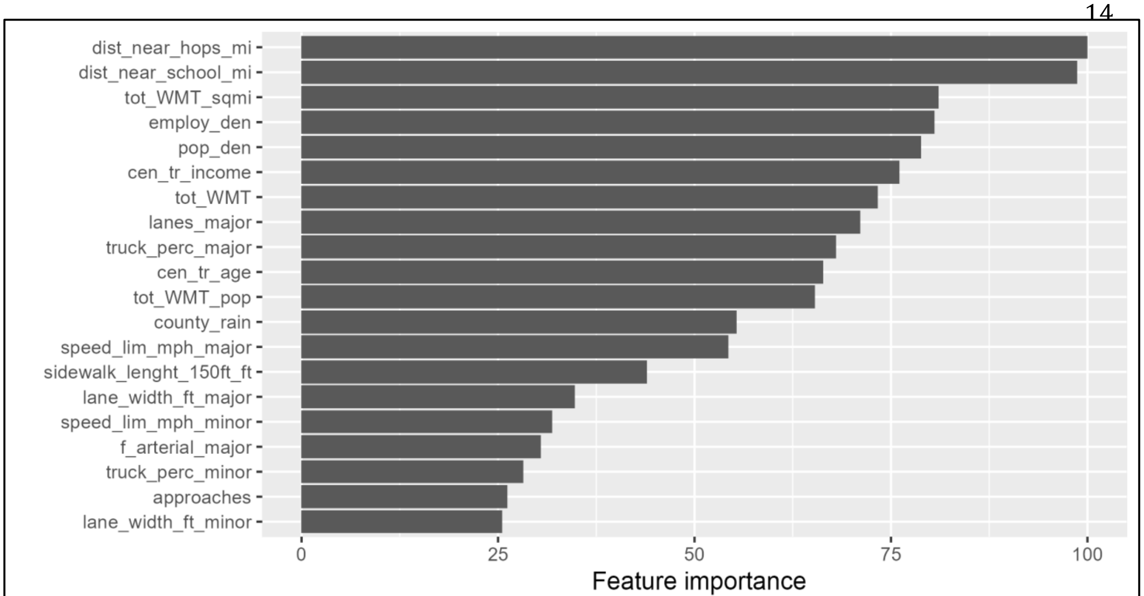


(b) Feature importance for balanced data

Figure 5: Feature importance for the RF models



(a) Signalized intersections



(b) Unsignalized intersections

Figure 6: Feature importance for the RF models for signalized and unsignalized intersections with sum (AADTs) ≥ 500

SENSITIVITY ANALYSIS OF CRASH PREDICTION

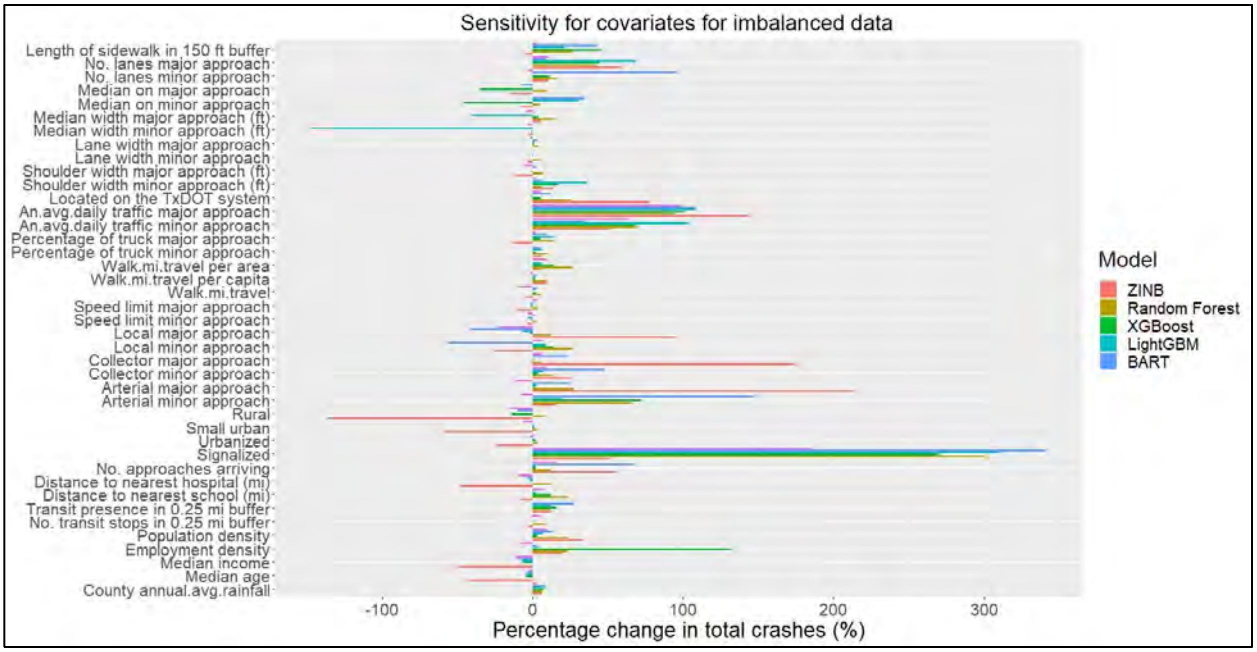
While regression and ML models excel at capturing relationships between features and outcome variables, the results may not be easy to interpret particularly for ML models. Specifically, one may find it difficult to quantify the substantive effects of each feature. Following Li and Kockelman (2022), this paper employed a sensitivity analysis that captured the contribution each variable had on the model's predictions. Let X be the set of features. The procedures of evaluating the sensitivity of variable X_i were as follows: (1) train the model on X and compute y as the prediction vector; (2) generate a new set X* where a transformation is performed on variable X_i; (3) generate prediction on X* and define $y*$ as the prediction vector, and (4) compute the percentage change in the prediction mean, denoted as $\frac{\overline{y*}-\bar{y}}{\bar{y}} * 100\%$ (Li and Kockelman 2022). Following Zuniga-Garcia, Perrine, and Kockelman (2022), the transformation was as follows: (1) increase one standard deviation for continuous features; (2) binary change (0 to 1; or 1 to 0) for dichotomous features. Essentially, one standard deviation or binary change was implemented on each observation (Rahman, Kockelman, and Perrine 2022). The new prediction was computed using the modified variables, and the difference between the mean of new predictions and original predictions represented the contribution of each feature (Zuniga-Garcia, Perrine, and Kockelman 2022).

This paper illustrates the sensitivities of each X_i by computing percentage changes in the outcome after performing transformation on each X_i. Specifically, we computed the percentage changes in the outcome variable, averaged across all 707,161 intersections, after one standard deviation change or binary change in each X_i. The higher the percentage changes, the higher contribution of a given variable on the model's predictions.

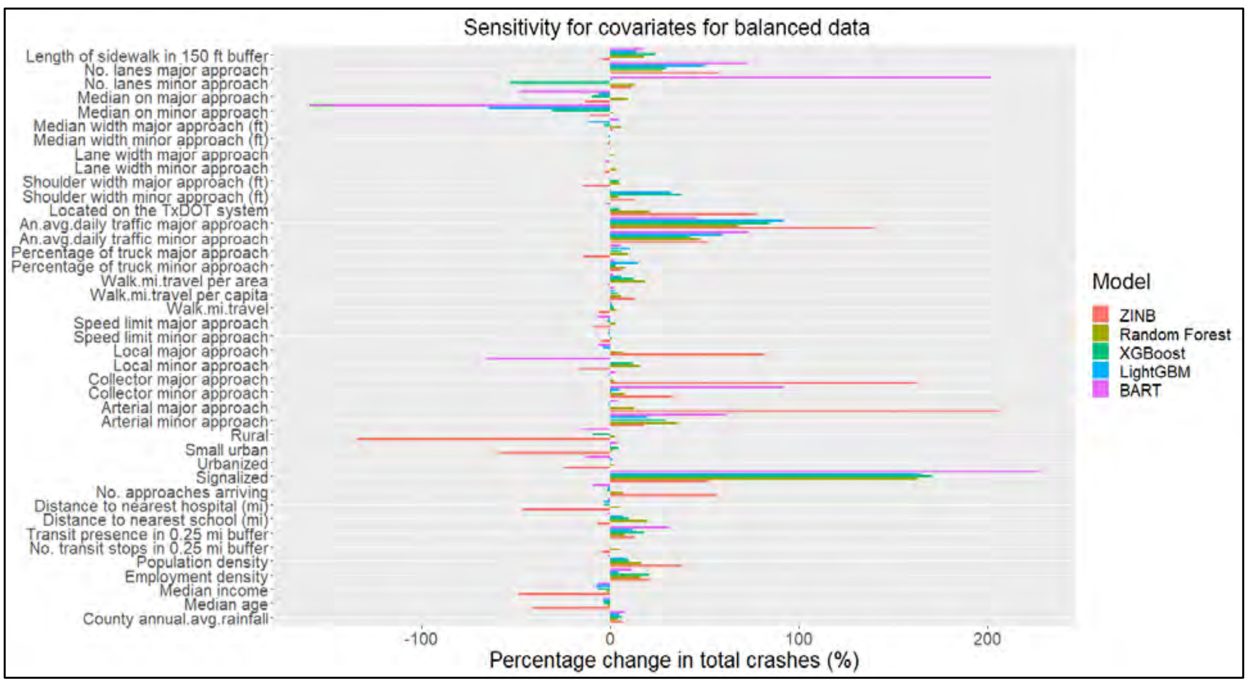
Figure 7 illustrates the sensitivities for the ZINB models and ML models for imbalanced and balanced data, respectively. One can see that the effects of several variables show different directions across different models. Considering the more important features, number of lanes at the minor approach, speed limits at the major and minor approaches, and distance to the nearest hospital show different directions in Figures 7a and 7b. This was possibly due to the fact that different ML models interpreted the significance of the features differently (Rahman, Kockelman, and Perrine 2022). Therefore, it is vital that one chooses the best performing model when one evaluates the metrics and examines feature importance with the optimum model. Since the ZINB model offers significance test of the predictor variables compared to ML models, this paper placed more weight on the results of the ZINB model when drawing inferences.

For the ZINB model, road types (local, collector, and arterial approaches) increased the outcome by a large percentage. In particular, arterial major approach had the most significant impact on the total number of crashes. A binary change on arterial major approach could lead to a 214% increase in crash occurrences per intersection. The percentage changes for land use characteristics were smaller in the ML models. For example, a binary change on arterial major approach resulted in less than a 30% increase in crash counts for all ML models. For ZINB models, intersections in rural areas, small urban, and urbanized areas decreased crash counts by 137%, 59%, and 24%, respectively, compared to large urban areas. In the ML models, the percentage changes pointed to different directions for different urban-rural classifications. For XGBoost and BART models, rural areas decreased crash counts while for the RF model, rural areas increased crash occurrences. Small urban and urbanized areas increased crash occurrences for most ML models. This contrasted with the ZINB results.

Concerning road design variables, the number of lanes and AADTs at the major and minor approaches had a significant impact on crash occurrence in the ZINB models. In the ZINB model, one standard deviation increase in the number of lanes at major approach led to a 59% increase in crash counts. In the RF model, the percentage change decreased to 42%. In the ZINB model, one standard deviation increase in AADTs at major and minor approaches contributed to about a 144% and 52% increase in crash occurrences, respectively. In the ML models, AADTs at major and minor approaches also showed increases in crash counts. The percentage increases ranged from 35% to 108% on imbalanced data, and from 42% to 92% on balanced data. Signalized intersections also contributed to a large increase in the outcome for both ZINB and ML models. A binary change on signalized intersections contributed to 300% and 163% increases in crash counts in the RF model on imbalanced and balanced data, respectively. As for census-tract level attributes, one standard deviation increase in population density, employment density, and precipitation increased crash counts by 33%, 20%, and 6%, respectively, while one standard deviation increase in median income and median age reduced crash occurrence by 50% and 43%, respectively. ML models showed the same directions in terms of percentage changes.

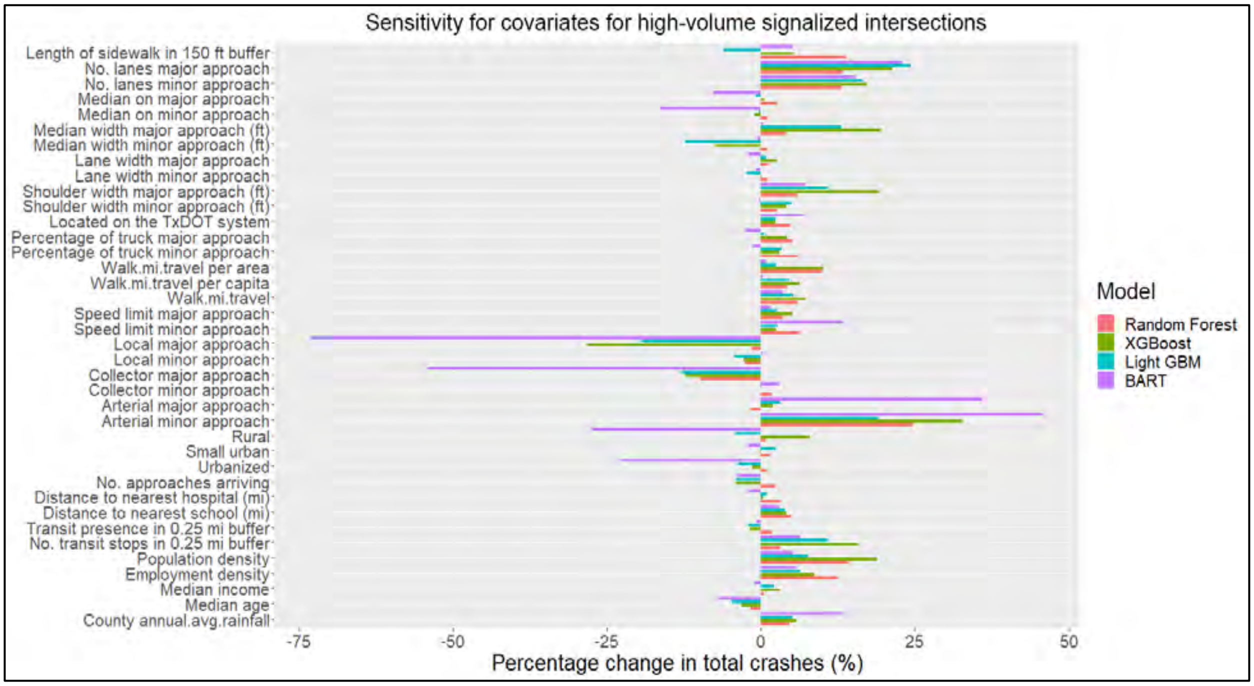


(a) Sensitivity for covariates for imbalanced data

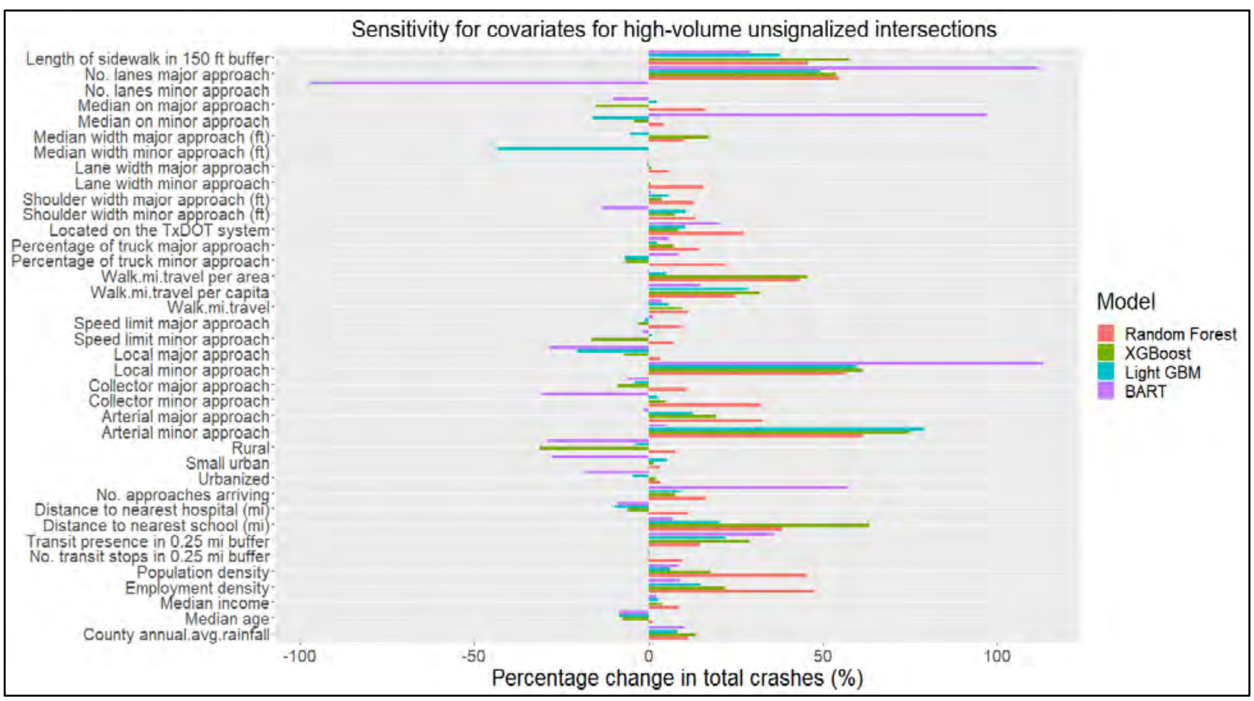


(b) Sensitivity for covariates for balanced data

Figure 7: Sensitivity analysis for covariates in predicting total crash occurrence



(a) Sensitivity for covariates for high-volume signalized intersections



(b) Sensitivity for covariates for high-volume unsignalized intersections

Figure 8: Sensitivity analysis of covariates in predicting total crash occurrences for high-volume (sum (AADT) ≥ 500) signalized and unsignalized intersections

This paper compared the sensitivity analysis results of the ML models for signalized and unsignalized intersections in Figures 8a and 8b, respectively. Comparing the two figures, most features showed similar directions in percentage changes. Focusing on the most important features, we found that the percentage changes for those variables were mostly consistent across the ML models. Consider the top five features. When one standard deviation was increased to total walk-miles traveled, distance to the nearest school, population density, and employment density one at a time, this paper found positive percentage changes in crash occurrences for all models. The only exception was distance to the nearest hospital. Among signalized intersections, all models showed positive percentage changes except BART, whereas among unsignalized intersections, all models demonstrated negative percentage changes except RF. It is noteworthy that RF yielded relatively large percentage changes for the top five features.

CONCLUSION

This study presented a comparison between five different models- one econometric and four ML models- to explore the opportunity for ML models in predicting the motor-vehicle crash frequency and injury counts at the intersections across Texas. R-square and RMSE metrics were used to evaluate the model fit and compare the model performances. Resampling of the data led to better prediction performances of all the models tested here and hence, the final comparison is made based on their performances on the balanced data. The ZINB model was found to be the least accurate model in terms of both R-square (-5.979) and RMSE (51.67). All four ML models provided much higher prediction accuracy than the ZINB model. The RF model offered the highest prediction accuracy among the ML models with R-square value of 0.832 and RMSE value of 8.59 for the balanced data. BART model had the lowest prediction accuracy among the ML models with R-square 0.602 and RMSE 27.53 followed by LightGBM with R-square 0.647 and RMSE 12.56. Though resampling increased the prediction accuracy for all the models, RF model saw the most significant improvement.

Employing the ML models to investigate the contributing factors of crash occurrence, this study found that signalized intersections and AADT both at minor and major approaches exerted disproportionately high weights on the model predictions. To deal with the problem, this paper subsetted the data into signalized intersections and unsignalized intersections and considered only the intersections where the sum of AADTs of the incoming links exceeded 500. Both for signalized and unsignalized intersections, RF model provided the highest accuracy in terms of both R-square (0.245 and 0.287) and RMSE (63.12 and 10.47) values. Analysis of the relative feature importance of the RF model for high-volume intersections (AADT > 500) showed that total walk-miles traveled, distance to the nearest school, distance to the nearest hospital, population density and employment density were the most important features to predict crash occurrence. Other important features included the number of lanes of the approaches, speed limit of the minor approach, arterial minor approach, and median income of the census tract.

Besides, the study carried out a sensitivity analysis to investigate traffic crash contributing factors by implementing one standard deviation increase (continuous features) and binary change (dichotomous features) for each observation. Sensitivity analysis showed that the effects of several variables have different directions across different models making interpreting their contribution in predicting crash occurrences difficult. Since the ZINB model offers significance test of the predictor variables compared to ML models, this paper placed more weight on the results of the ZINB model when drawing inferences. The ZINB model showed that road types of the approaches (local, collector, and arterial approaches) increase crash frequency by a large percentage (214%) compared to the ML models. On the other hand, a binary change on the arterial major approach resulted in less than a 30% increase in crash counts for all ML models. For ZINB models, intersections in rural areas, small urban, and urbanized areas decreased crash counts by 137%, 59%, and 24%, respectively, compared to large urban areas. The percentage changes for different urban-rural classifications showed different directions in the ML models. For XGBoost and BART models, rural areas decreased crash counts while for the RF model, rural areas increased crash occurrences. This made interpreting the influence of the predictors difficult and unreliable for the ML models. Among the road design variables, one standard deviation increase in the number of lanes and AADTs at the major and minor approaches significantly increases crash count in the ZINB model. In the ML models, an increase of AADTs also increased crash count, but an increase in the number of lanes led to a decrease in crash count, contrasting the findings from ZINB model. Signalized intersections had been found to increase the crash count both in the ZINB and ML models. Among the census-tract level predictors, an increase in population density, employment density, and precipitation increased crash counts whereas the increase in median income and median age reduced crash occurrences. Both ZINB and ML models showed similar directions for the census-tract level variables.

Summing up, this paper concurs with other similar studies (Iranitalab and Khattak 2017; Rahim and Hassan 2021) upon the fact that ML models are better at predicting crash occurrences whereas statistical models are better at investigating the contributing factors of a crash event. The lack of test of significance and fluctuations of sensitivity of the predictor variables across models make the result ambiguous and unreliable. Different settings of the ML models may provide different results and change the drawn inferences. Traffic and transportation agencies can use ML models in predicting a crash event with higher accuracy, but care should be taken while investigating pre-crash conditions and influencing factors using ML models.

REFERENCES

Abdelwahab, H. T., and M. A. Abdel-Aty. 2001. "Development of Artificial Neural Network Models to Predict Driver Injury Severity in Traffic Accidents at Signalized Intersections." *Transportation Research Record* 1746 (1): 6-13. doi: https://doi.org/10.3141/1746-02.

ACS (2020). "American Community Survey Data via API." Accessed 1 September 2022. https://www.census.gov/programs-surveys/acs/data/data-via-api.html. See codes from the "get_census_tract.R" replication file.

CDC. (2018). WISQARS (Web-based Injury Statistics Query and Reporting System). Atlanta, GA: US Department of Health and Human Services. Accessed 4 January 2018. https://www.cdc.gov/injury/wisqars.

Caliendo, C., M. Guida, and A. Parisi. 2007. "A Crash-prediction Model for Multilane Roads." *Accident Analysis & Prevention* 39(4): 657-670. doi: https://doi.org/10.1016/j.aap.2006.10.012.

Casalicchio, G., C. Molnar, and B. Bischl. 2018. "Visualizing the Feature Importance for Black Box Models." *ECML PKDD 2018*: 655-670. doi: https://doi.org/10.48550/arXiv.1804.06620.

Chang, L., and W. Chen. 2005. "Data Mining of Tree-based Models to Analyze Freeway Accident Frequency." *Journal of Safety Research* 36: 365–375. doi: https://doi.org/10.1016/j.jsr.2005.06.013.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*: 785-794. doi: https://doi.org/10.48550/arXiv.1603.02754.

Chipman, H. A., E. I. George, and R. E. McCulloch. 2010. "Bart: Bayesian Additive Regression Trees." *The Annals of Applied Statistics* 4 (1): 266-298. doi: https://doi.org/10.1214/09-AOAS285.

Cho, K., B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio. 2014. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*: 1724-1734. doi: https://doi.org/10.48550/arXiv.1406.1078.

Chong, M., A. Abraham, and M. Paprzycki. 2005. "Traffic Accident Analysis Using Machine Learning Paradigms." *Informatica* 29: 89-98.

C.R.I.S., Texas Department of Transportation (2020). C.R.I.S. Query. Accessed 1 September 2022. https://cris.dot.state.tx.us/public/Query/app/home.

de Ona, J., G. López, R. Mujalli, and F. J. Calvo. 2013. "Analysis of Traffic Accidents on Rural Highways Using Latent Class Clustering and Bayesian Networks." *Accident Analysis & Prevention* 51: 1-10. doi: https://doi.org/10.1016/j.aap.2012.10.016.

Dong, C., D. B. Clarke, X. Yan, A. Khattak, and B. Huang. 2014. "Multivariate Random-parameters Zero-inflated Negative Binomial Regression Model: An Application to Estimate Crash Frequencies at Intersections." *Accident Analysis & Prevention* 70: 320-329. doi: https://doi.org/10.1016/j.aap.2014.04.018.

Dong, C., C. Shao, J. Li, and Z. Xiong. 2018. "An Improved Deep Learning Model for Traffic Crash Prediction." *Journal of Advanced Transportation*, *2018*. doi: https://doi.org/10.1155/2018/3869106.

Dionne, G., D. Desjardins, C. Laberge-Nadeau, and U. Maag. 1995. "Medical Conditions, Risk Exposure, and Truck Drivers' Accidents: An Analysis with Count Data Regression Models." *Accident Analysis & Prevention* 27 (3): 295-305. doi: https://doi.org/10.1016/0001-4575(94)00071-s.

Elamrani Abou Elassad, Z., H. Mousannif, and H. Al Moatassime. 2020. "Class-imbalanced Crash Prediction Based on Real-time Traffic and Weather Data: A Driving Simulator Study." *Traffic Injury Prevention* 21 (3): 201-208. doi: https://doi.org/10.1080/15389588.2020.1723794.

Eluru, N., M. Bagheri, L. F. Miranda-Moreno, and L. Fu. 2012. "A Latent Class Modeling Approach for Identifying Vehicle Driver Injury Severity Factors at Highway-railway Crossings." *Accident Analysis & Prevention* 47: 119-127. doi: https://doi.org/10.1016/j.aap.2012.01.027.

Fiorentini, N., and M. Losa. 2020. "Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms." *Infrastructures* 5 (7), 61. doi: https://doi.org/10.3390/infrastructures5070061.

Greenwell, B. M., and B. C. Boehmke. 2020. "Variable Importance Plots— An Introduction to the vip Package." *The R Journal* 12 (1): 343-366. doi: https://doi.org/10.32614/RJ-2020-013.

Guo, X., Y. Yin, C. Dong, G. Yang, and G. Zhou. 2008. "On the Class Imbalance Problem." In *2008 Fourth International Conference on Natural Computation* (Vol. 4, pp. 192-201). IEEE. doi: https://doi.org/10.1109/ICNC.2008.871.

Greene, W. 2007. "Functional Form and Heterogeneity in Models for Count Data." *Foundations and Trends in Econometrics* 1 (2): 113-218. doi: https://doi.org/10.1561/0800000008.

He, H., and E. A. Garcia. 2009. "Learning from Imbalanced Data." *IEEE Transactions on Knowledge and Data Engineering* 21 (9): 1263-1284. doi: https://doi.org/10.1109/TKDE.2008.239.

Iranitalab, A., and A. Khattak. 2017. "Comparison of Four Statistical and Machine Learning Methods for Crash Severity Prediction." *Accident Analysis & Prevention* 108: 27-36. doi: https://doi.org/10.1016/j.aap.2017.08.008.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. New York, NY: Springer.

Jovanis, P. P., and H. Chang. 1986. "Modeling the Relationship of Accidents to Miles Traveled." *Transportation Research Record* 1068: 42-51.

Kaplan, S., and C. G. Prato. 2013. "Cyclist–motorist Crash Patterns in Denmark: A Latent Class Clustering Approach." *Traffic Injury Prevention* 14 (7): 725-733. doi: https://doi.org/10.1080/15389588.2012.759654.

Karlaftis, M. G., and E. I. Vlahogianni. 2011. "Statistical Methods versus Neural Networks in Transportation Research: Differences, Similarities and Some Insights." *Transportation Research Part C: Emerging Technologies* 19 (3): 387-399. doi: https://doi.org/10.1016/j.trc.2010.10.004.

Ke, G., Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T. Liu. 2017. "LightGBM: A Highly Efficient Gradient Boosting Decision Tree." *Proceedings of the 31st Conference on Neural Information Processing Systems*: 3149-3157.

Kingma, D. P., and J. Ba. 2017. "Adam: A Method for Stochastic Optimization." *Proceedings of the 3rd International Conference for Learning Representations, San Diego, 2015*. doi: https://doi.org/10.48550/arXiv.1412.6980.

Khattak, A. J., P. Kantor, and F. M. Council. 1998. "Role of Adverse Weather in Key Crash Types on Limited-access: Roadways Implications for Advanced Weather Systems." *Transportation Research Record* 1621 (1): 10-19. doi: https://doi.org/10.3141/1621-02.

Khattak, A. J., M. D. Pawlovich, R. R. Souleyrette, and S. L. Hallmark. 2002. "Factors Related to More Severe Older Driver Traffic Crash Injuries." *Journal of Transportation Engineering* 128 (3): 243-249.

Li, W., and K. M. Kockelman. 2022. "How does Machine Learning Compare to Conventional Econometrics for Transport Data Sets? A Test of ML versus MLE." *Growth and Change* 53 (1): 342–376. doi: https://doi.org/10.1111/grow.12587.

Liaw, A., and M. Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22.

Liu, P., S. Chen, and M. Yang. 2008. "Study of Signalized Intersection Crashes Using Artificial Intelligence Methods." In *MICAI 2008: Advances in Artificial Intelligence. MICAI 2008. Lecture Notes in Computer Science* 5317, edited by A. Gelbukh, and E. F. Morales. Berlin: Springer. doi: https://doi.org/10.1007/978-3-540-88636-5_93.

Lord, D., and F. Mannering. 2010. "The Statistical Analysis of Crash-frequency Data: A Review and Assessment of Methodological Alternatives." *Transportation Research Part A: Policy and Practice* 44 (5): 291-305. doi: https://doi.org/10.1016/j.tra.2010.02.001.

Lunardon, N., G. Menardi, and N. Torelli. 2014. "ROSE: a Package for Binary Imbalanced Learning." *The R Journal* 6 (1): 79-89. doi: https://doi.org/10.32614/RJ-2014-008.

National Center for Statistics and Analysis. (2017, October). *2016 Fatal Motor Vehicle Crashes: Overviewexternal Icon*. (Traffic Safety Facts Research Note. Report No. DOT HS 812 456). Washington, DC: National Highway Traffic Safety Administration.

Rahim, M. A., and H. M. Hassan. 2021. "A Deep Learning Based Traffic Crash Severity Prediction Framework." *Accident Analysis & Prevention* 154, 106090. doi: https://doi.org/10.1016/j.aap.2021.106090.

Rahman, M. S., M. Abdel-Aty, S. Hasan, and Q. Cai. 2019. "Applying Machine Learning Approaches to Analyze the Vulnerable Road-users' Crashes at Statewide Traffic Analysis Zones." *Journal of Safety Research* 70: 275-288. doi: https://doi.org/10.1016/j.jsr.2019.04.008.

Rahman, M., K. M. Kockelman, and K. A. Perrine. 2022. "Investigating Risk Factors Associated with Pedestrian Crash Occurrence and Injury Severity in Texas." *Traffic Injury Prevention* 23 (5): 283-289. doi: https://doi.org/10.1080/15389588.2022.2059474.

Texas Water Development Board. (2014). GIS Data: Texas Precipitation. Accessed 10 December 2021. https://www.twdb.texas.gov/mapping/gisdata.asp.

UCLA. 2022. Negative Binomial Regression: Stata Annotated Output. UCLA: Statistical Consulting Group. Accessed 10 December 2021. https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-how-do-i-cite-web-pages-and-programs-from-the-ucla-statistical-consulting-group/.

van der Laan, M. J. 2006. “Statistical Inference for Variable Importance.” *The International Journal of Biostatistics* 2 (1): 1-33. doi: https://doi.org/10.2202/1557-4679.1008.

Wang, Y., C. M. Monsere, C. Chen, and H. Wang. 2018. “Development of a Crash Risk-Scoring Tool for Pedestrian and Bicycle Projects in Oregon.” *Transportation Research Record* 2672 (32): 30-39. doi: https://doi.org/10.1177/0361198118794285.

Yasmin, S., and N. Eluru. 2018. “A Joint Econometric Framework for Modeling Crash Counts by Severity.” *Transportmetrica A Transport Science* 14 (3): 230-255. doi: https://doi.org/10.1080/23249935.2017.1369469.

Yu, R., and M. Abdel-Aty. 2013. “Utilizing Support Vector Machine in Real-time Crash Risk Evaluation.” *Accident Analysis & Prevention* 51: 252-259. doi: https://doi.org/10.1016/j.aap.2012.11.027.

Zhao, S., A. Iranitalab, and A. J. Khattak. 2019. “A Clustering Approach to Injury Severity in Pedestrian-train Crashes at Highway-rail Grade Crossings.” *Journal of Transportation Safety & Security* 11 (3): 305-322. doi: https://doi.org/10.1080/19439962.2018.1428257.

Zheng, M., T. Li, R. Zhu, J. Chen, Z. Ma, M. Tang, Z. Cui, and Z. Wang. 2019. “Traffic Accident’s Severity Prediction: A Deep-Learning Approach-Based CNN Network.” *IEEE Access* 7: 39897-39910. doi: https://doi.org/10.1109/ACCESS.2019.2903319.

Zhao, B., N. Zuniga-Garcia, L. Xing, and K. M. Kockelman. 2021. “Predicting Pedestrian Crash Occurrence and Injury Severity in Texas Using Tree-based Machine Learning Models.” *Under Review for publication in Traffic Injury Prevention*.

Zuniga-Garcia, N., K. A. Perrine, and K. M. Kockelman. 2022. “Predicting Pedestrian Crashes in Texas’ Intersections and Midblock Segments.” *Sustainability* 14 (12), 7164. doi: https://doi.org/10.3390/su14127164.